

Homework #1

See Canvas for HW #1 assignment due date. Complete all of the following problems. Ideally, the theoretical problems should be answered in a Markdown cell directly underneath the question. If you don't know LaTeX/Markdown, you may submit separate handwritten solutions to the theoretical problems, but please see the class scanning policy. Please do not turn in messy work. Computational problems should be completed in this notebook (using the R kernel). Computational questions may require code, plots, analysis, interpretation, etc. Working in small groups is allowed, but it is important that you make an effort to master the material and hand in your own work.

A. Theoretical Problems

Problem A.1

Suppose that $X_1, \dots, X_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$, where σ is known, and we are interested in an estimator for $\theta = \mu^2$. (Note that we will use the following calculations to make comparisons to the parametric bootstrap method explored below).

- (a) Find the maximum likelihood estimator (MLE) for θ , denoted $\hat{\theta}$.
- (b) Compute the bias of $\hat{\theta}$, denoted $Bias(\hat{\theta})$. Recall that $Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$.
- (c) Compute the variance of $\hat{\theta}$, denoted $Var(\hat{\theta})$. (HINT: You might use a moment generating function at some point in your answer.)
- (d) Write down the bootstrap estimators of $Bias(\hat{\theta})$ and $Var(\hat{\theta})$.

Problem A.2

Provide a brief explanation of the pros and cons of using the bootstrap for calculating confidence intervals.

B. Computational Problems

Problem B.1

Suppose that $X_1, \dots, X_8 \stackrel{iid}{\sim} \Gamma(\alpha, \beta)$. Let's use the bootstrap to compute a 90% confidence interval for the population standard deviation: $sd(X) = \sqrt{\alpha/\beta^2} = \theta$.

Note:

The convention in this course will be to interpret $\Gamma(\alpha, \beta)$ as the "shape/rate" parameterization: shape = α , rate = β . But R uses the "shape/scale" parameterization: shape = α , scale = $\theta = 1/\beta$.

To be sure that you are properly simulating from the right gamma distribution, see the help file for `rgamma()` (run: ?
`rgamma`). Also, see [here](#) for more information on the gamma distribution.

(a) State why a χ^2 confidence interval is not valid in this context.

You should reply on knowledge from your prereq class!

(b) Generate a sample of size $n = 8$ from $\Gamma(\alpha = 3, \beta = 4)$ and calculate the true population standard deviation (in this example, we are generating data so that we can see how well our estimation procedure will do).

(c) Generate $B = 200$ bootstrap samples from the above sample. Print the dimension, and articulate what each row/column represents. Avoid loops! (HINT: use the `replicate()` function.)

(d) Calculate and print the sample standard deviation, s . Then, calculate s for each bootstrap sample. Denote this as s_i^* , for $i = 1, \dots, B$. Avoid loops! (HINT: use the `apply()` function.) Display a histogram of the distribution of s_i^* , $i = 1, \dots, B$.

(e) Use the `quantile()` function to find the 5th and 95th percentile of the distribution of s_i^* . Use these values to calculate the 90% bootstrap pivot confidence interval and bootstrap percentile confidence interval for θ .

(f) Interpret this confidence interval.

Problem B.2

Thus far, we've been looking at the *nonparametric bootstrap*. In this problem, we look at the *parametric bootstrap* as a way of estimating the bias and variance of an estimator $\hat{\theta} = \bar{X}^2$ of $\theta = \mu^2$ (in problem A.1 you calculated these values exactly).

(a) Generate $X_1, \dots, X_{20} \stackrel{iid}{\sim} N(\mu = 2, \sigma^2 = 1)$, and then forget that you know μ and σ^2 . Find the sample mean and sample variance.

(b) Define \widehat{N} to be the distribution of the variable X_i in the population with the sample estimates plugged in for the unknown population parameters. Write down \widehat{N} based on the data generated in (a).

(c) Draw $B = 500$ parametric bootstrap samples from \widehat{N} , and for each bootstrap sample $(X_{1,j}, \dots, X_{20,j})$, compute

$$\hat{\theta}_j^* = \left(\frac{1}{20} \sum_{i=1}^{20} X_{i,j}^* \right)^2,$$

where $j = 1, \dots, B$.

(d) Compute an estimate of the bias:

$$\widehat{B}(\hat{\theta}) \approx \frac{1}{B} \sum_{j=1}^B \hat{\theta}_j^* - \bar{x}^2$$

Compare this to the exact bias using the formula in problem A.1.

(e) Compute an estimate of the variance:

$$\widehat{\text{Var}}(\hat{\theta}) \approx \frac{1}{B-1} \sum_{j=1}^B (\hat{\theta}_j^* - \bar{\theta})^2,$$

where

$$\bar{\theta} = \frac{1}{B} \sum_{j=1}^B \hat{\theta}_j^*.$$

Compare this to the exact variance using the formula in problem A.1.

(f) True or False: For a fixed sample size $n = 20$, as B increases, $\widehat{\text{Bias}}(\hat{\theta})$ will approach $\text{Bias}(\hat{\theta})$. That is, for a fixed n , the bootstrap estimate of the bias will approach the true bias as the number of bootstrap samples, B increases. You might consider running a simulation to decide!

Problem B.3

The "Wisconsin Card Sorting Test" is widely used by psychiatrists, neurologists, and neuropsychologists with patients who have a brain injury. Patients with any sort of frontal lobe lesion generally do poorly on the test. The data frame `WCST` contains the test scores from a group of 50 patients from the *Virgen del Camino* Hospital.

(a) Using the code below, load the `WCST` data and explore whether there is reason to believe that the score data comes from a non-normal distribution. First, create a histogram (use `ggplot!`) and describe whether the data look normal. Then, use the function `shapiro.test()` to explore normality. Be sure to explain what this function does--i.e., what's the null and alternative hypothesis--in your answer.

(b) What assumptions must be made in order to compute a (non-bootstrap) 95% confidence interval for the population mean score?

(c) Compute the confidence interval from (b).

(d) Compute a 95% bootstrap pivot confidence interval for the mean.

Problem B.4

The dataset gives the number of births per month in New York city, from January 1946 to December 1959. The data are ordered.

(a) Construct another column in the dataset that labels the month and year for each birth per month record.

(b) Construct a plot of births per month against the month/year column that you created in part (a). Analyze the plot. Do you notice anything interesting?

(c) Suppose that your boss asked you to use the bootstrap to construct a confidence interval for the average number of births per month in New York city over the time period in the dataset. Write a short response to your boss describing why this confidence interval is not valid for this data.